BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 10: e182512
doi: 10.3897/biss.10.182512

OPEN ACCESS

Conference Abstract

Multimodal Question Answering for Botanical Archives: Adapting PaLI-GEMMA 2 to Enhance Exploration of Herbarium Collections

Ahmed Zine Eddine Marhoum‡, Anis Oukid‡, Hanane Ariouat§, Nicolas Turenne§, Eric Chenin§, Youcef Sklab§

‡ ESTIN, Bejaia, Algeria
§ IRD, Sorbonne Université, UMMISCO, Bondy, France

Corresponding author: Youcef Sklab (youcef.sklab@ird.fr)

Received: 15 Dec 2025 | Published: 09 Jan 2026

Citation: Marhoum AZE, Oukid A, Ariouat H, Turenne N, Chenin E, Sklab Y (2026) Multimodal Question Answering for Botanical Archives: Adapting PaLI-GEMMA 2 to Enhance Exploration of Herbarium Collections. Biodiversity Information Science and Standards 10: e182512. https://doi.org/10.3897/biss.10.182512

Abstract

Herbarium collections serve as irreplaceable archives for taxonomic, ecological, and evolutionary research. Recent advances in deep learning now make it possible to design AI models capable of automatically annotating herbarium specimens (Ariouat et al. 2025, Sklab et al. 2025a, Sklab et al. 2025b), to develop intelligent systems for querying, visualizing, and exploring them. In previous work Sklab et al. 2025a, state-of-the-art Computer Vision (CV) models failed to generalize to unseen specimens, relying on superficial cues like labels and mounting paper rather than learning true plant morphology. Meanwhile, rich botanical texts remain disconnected from specimen images: CV models lack semantic depth, and LLMs extract traits but can't ground them visually.

Vision–Language Models (VLMs) emerge as a natural solution, combining the strengths of both to close this gap. By jointly reasoning across images and text, they have the potential to align visual perception with botanical semantics. However, most existing VLMs are trained on general-purpose datasets and fail to capture the specialized features inherent to herbarium specimens. To address this limitation, we investigate how

© Marhoum A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

adapting a VLM to the botanical domain enables it to answer expert-level questions about individual herbarium specimens.

We adapted the **PaliGemma-2** model (Andreas 2024), leveraging its Visual Question Answering (VQA) capabilities to align large language model knowledge with the observable content of herbarium images. In this setup, the digitized specimen is provided as visual context, and the user interacts with the model through targeted questions. The adapted model grounds its answers in visible features of the specimen—morphology, phenology, and context—rather than relying on generic species-level knowledge. By doing so, the system aims to exploit the semantic richness encoded in modern large language models while constraining it to the biological reality captured in a single herbarium sheet — a snapshot of the plant at the moment it was collected, rather than an idealized description of its entire life history. Our approach comprises three stages (Fig. 1): i) domain-specific dataset construction; ii) parameter-efficient fine-tuning; and iii) evaluation.

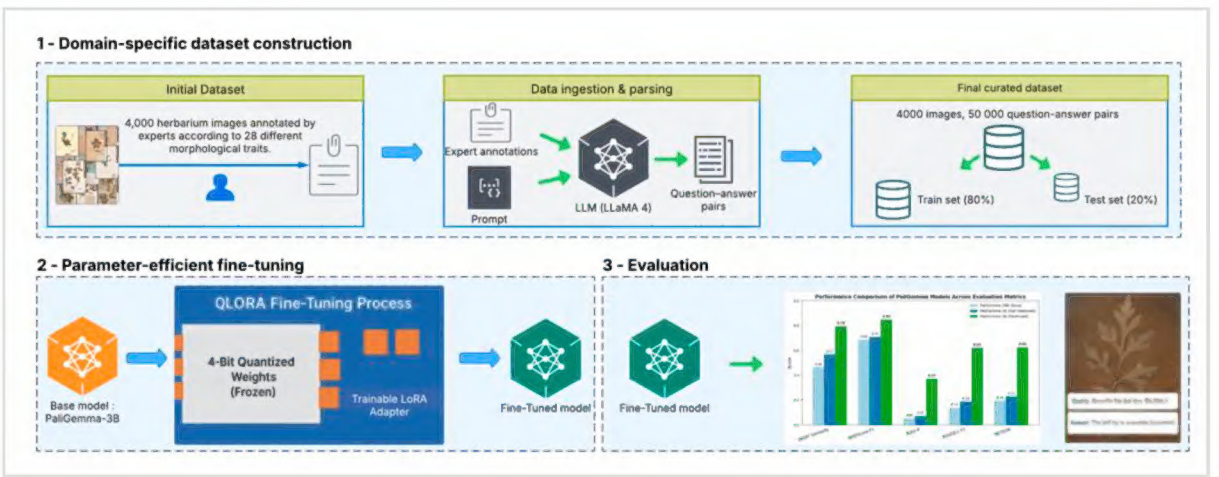

Figure 1.

Pipeline for adapting a vision–language model to herbarium specimens through expert annotations, automated question–answer generation, and fine-tuning. Credit: Marhoum et al., © 2025 The Authors. Licensed under CC BY 4.0.

We curated a dataset from herbarium specimens, combining digitized images with expert-annotations (28 traits). To translate structured trait annotations into a linguistically and semantically rich training corpus, we employed the **LLaMA 4** (Meta AI 2025) language model to automatically generate over 50,000 question–answer pairs grounded in botanical semantics. To achieve domain adaptation under limited computational resources, we employed **Quantization-aware Low-Rank Adaptation (QLoRA)**, which enables efficient fine-tuning by introducing trainable low-rank adapter layers into a quantized pre-trained backbone. We evaluated the fine-tuned model against two baselines, its un-fine-tuned version and the larger **PaliGemma-28B** model, using five metrics (Fig. 2) to quantify both semantic similarity and lexical fidelity with respect to expert annotations. Fine-tuning yielded substantial improvements across all evaluation metrics. The adapted PaliGemma-3B achieved a 40% increase in SBERT similarity (Fig. 2) relative to its base counterpart, reflecting enhanced semantic understanding of botanical traits. Likewise, gains in ROUGE-L and METEOR demonstrate the model's

capacity to employ precise botanical terminology rather than generic descriptive language. This highlights the model's potential to ground linguistic outputs in visual evidence, enabling more accurate interpretation of herbarium specimens.

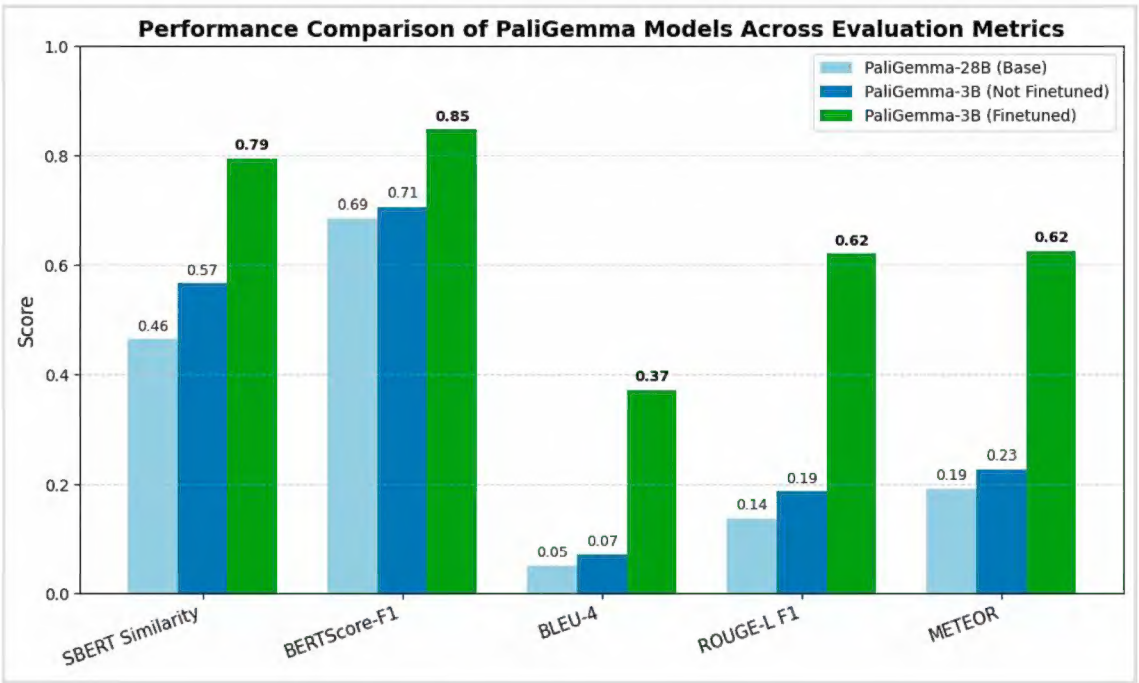

Figure 2.

Quantitative comparison demonstrating that the fine-tuned PaLI-GEMMA-3B model substantially outperforms both the unfine-tuned and the baseline model. Credit: Marhoum et al., © 2025 The Authors. Licensed under CC BY 4.0.

Keywords

plant trait detection, vision-language models, visual question answering

Presenting author

Ahmed Zine Eddine Marhoum

Presented at

Living Data 2025

Hosting institution

Institut de Recherche pour le Développement (IRD)

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Andreas S, et al. (2024) PaliGemma 2: A Family of Versatile VLMs for Transfer. arXiv URL: https://arxiv.org/abs/2407.07726
- Ariouat H, Sklab Y, Prifti E, Zucker J, Chenin E (2025) Enhancing plant morphological trait identification in herbarium collections through deep learning–based segmentation. Applications in Plant Sciences https://doi.org/10.1002/aps3.70000
- Meta AI (2025) LLaMA 4 Scout & Maverick model card. https://www.llama.com/models/llama-4/. Accessed on: 2025-12-23.
- Sklab Y, Ariouat H, Chenin E, Prifti E, Zucker J (2025a) SIM-Net: A Multimodal Fusion Network Using Inferred 3D Object Shape Point Clouds From RGB Images for 2D Classification. IET Computer Vision https://doi.org/10.1049/cvi2.70036
- Sklab Y, Castanet F, Ariouat H, Arib S, Zucker J, Chenin E, Prifti E (2025b) PlantSAM: An Object Detection-Driven Segmentation Pipeline for Herbarium Specimens. Applications in Plant Sciences https://doi.org/10.1002/aps3.70034